INFOMDA2: Battling the curse of dimensionality

Erik-Jan van Kesteren

8/10/2021

Course proceedings

Recap: INFOMDA1

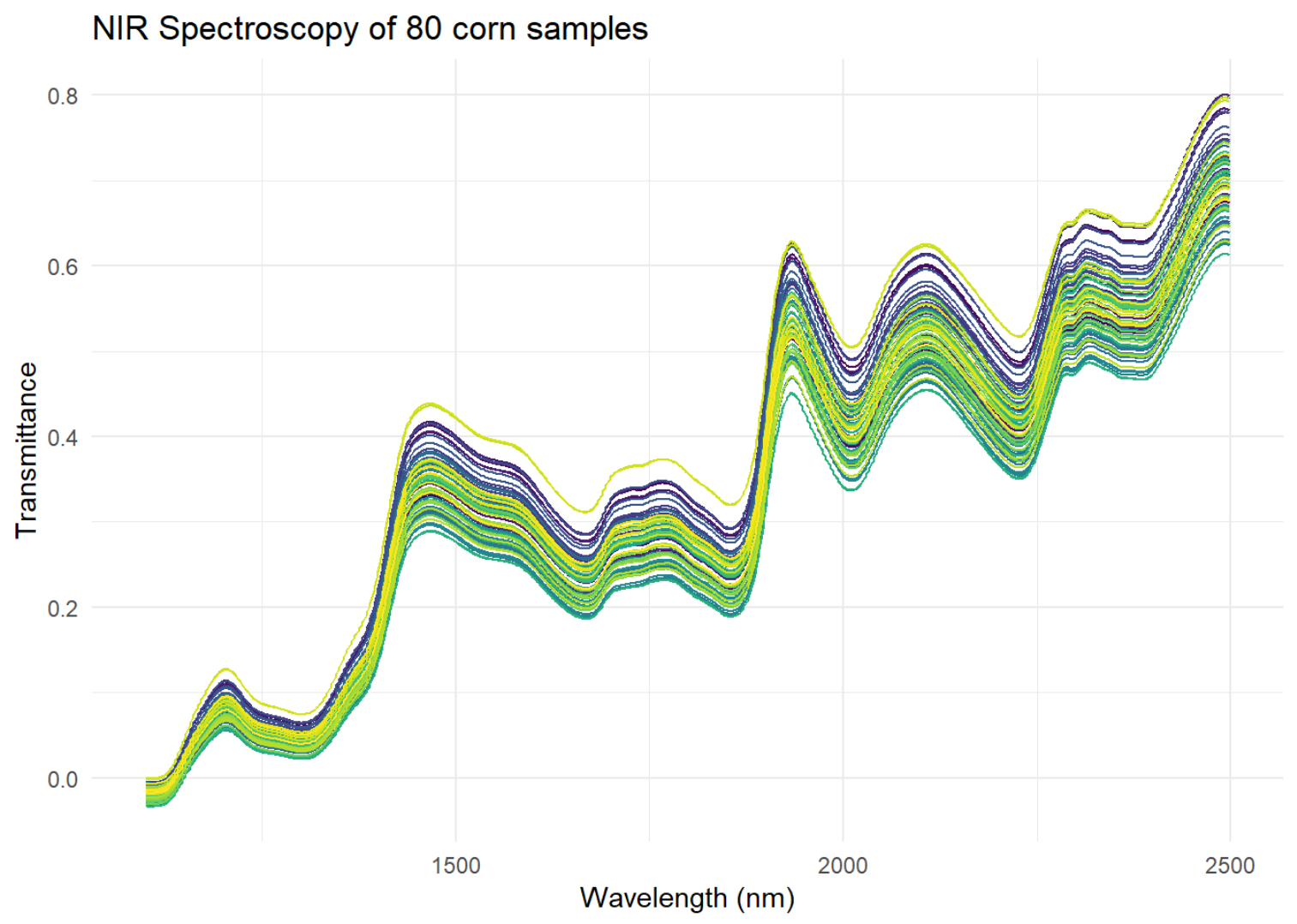
- Bullet 1
- Bullet 2
- Bullet 3

- Bullet 1
- Bullet 2
- Bullet 3

Theme: High-dimensionality

Example 1: Genomic data

Example 2: Spectroscopy of corn samples



Example 3: Deep Neural networks

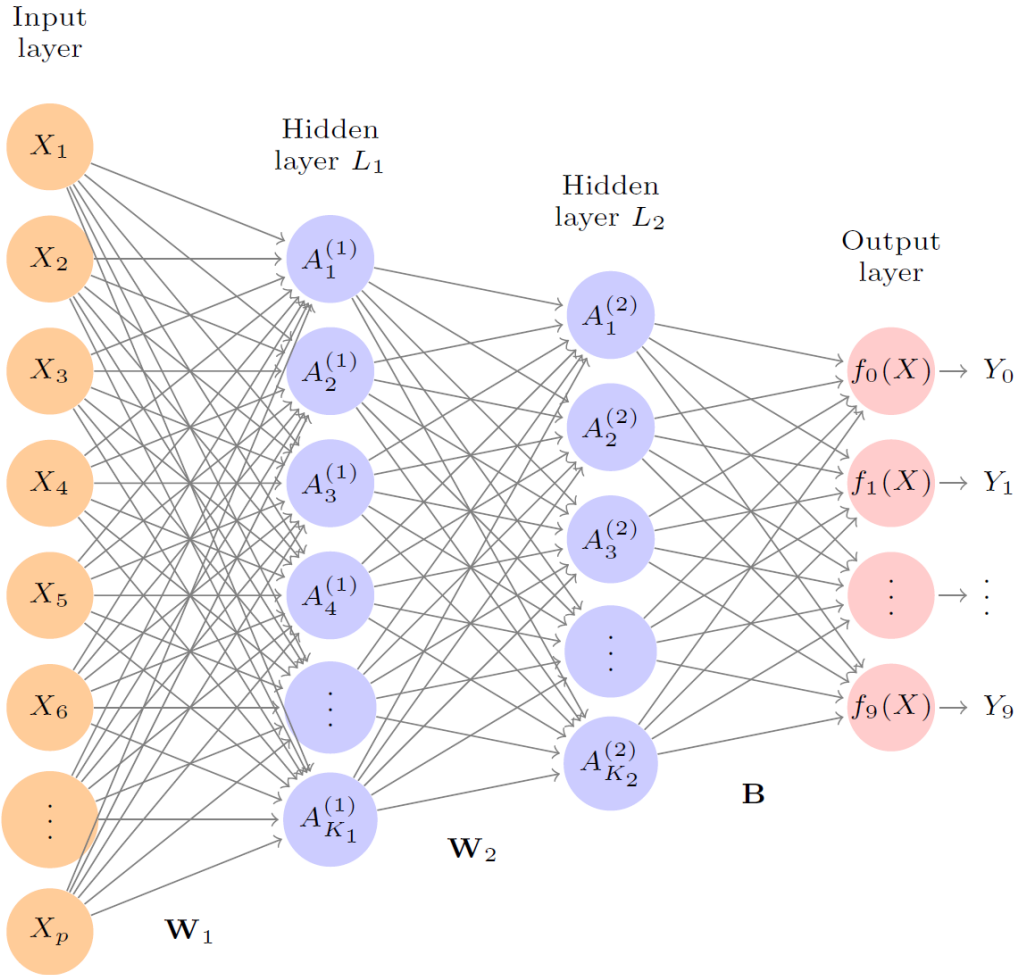


FIGURE 10.4. *Neural network diagram with two hidden layers and multiple*

What is high-dimensionality?

- Dataset with many columns and few rows
- Dataset with many transformed features and few observations
- Neural networks with many parameters

High-dimensional: many parameters (p) *relative to* the amount of information (N) available to learn about those parameters.

The curse of dimensionality

The curse of dimensionality

Let's do linear regression:

$$y_i = \beta_0 + \beta_1 \cdot x_i + \epsilon_i$$

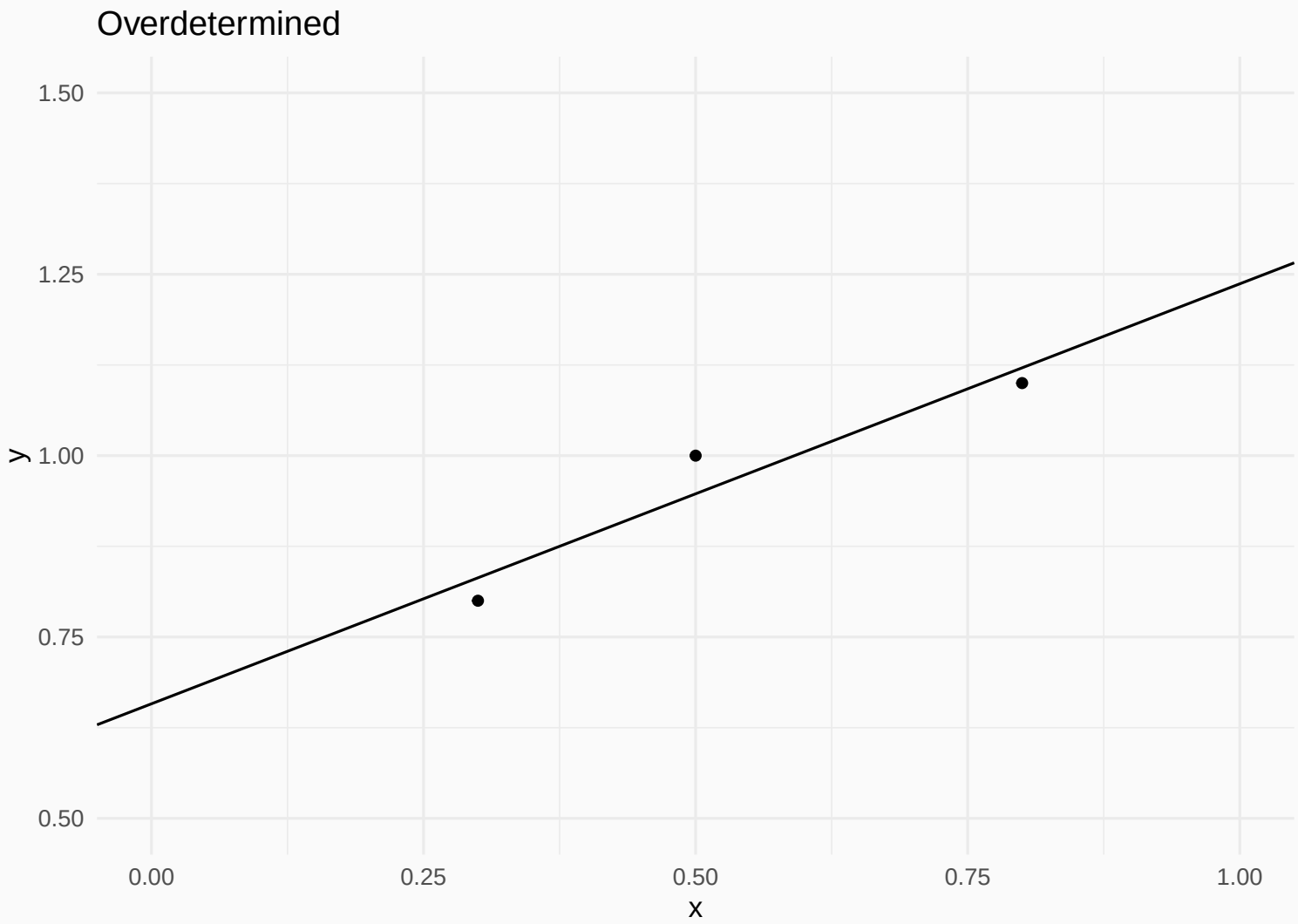
Dataset:

```
## # A tibble: 3 x 2
##       x     y
##   <dbl> <dbl>
## 1   0.5   1
## 2   0.3   0.8
## 3   0.8   1.1
```

Estimation:

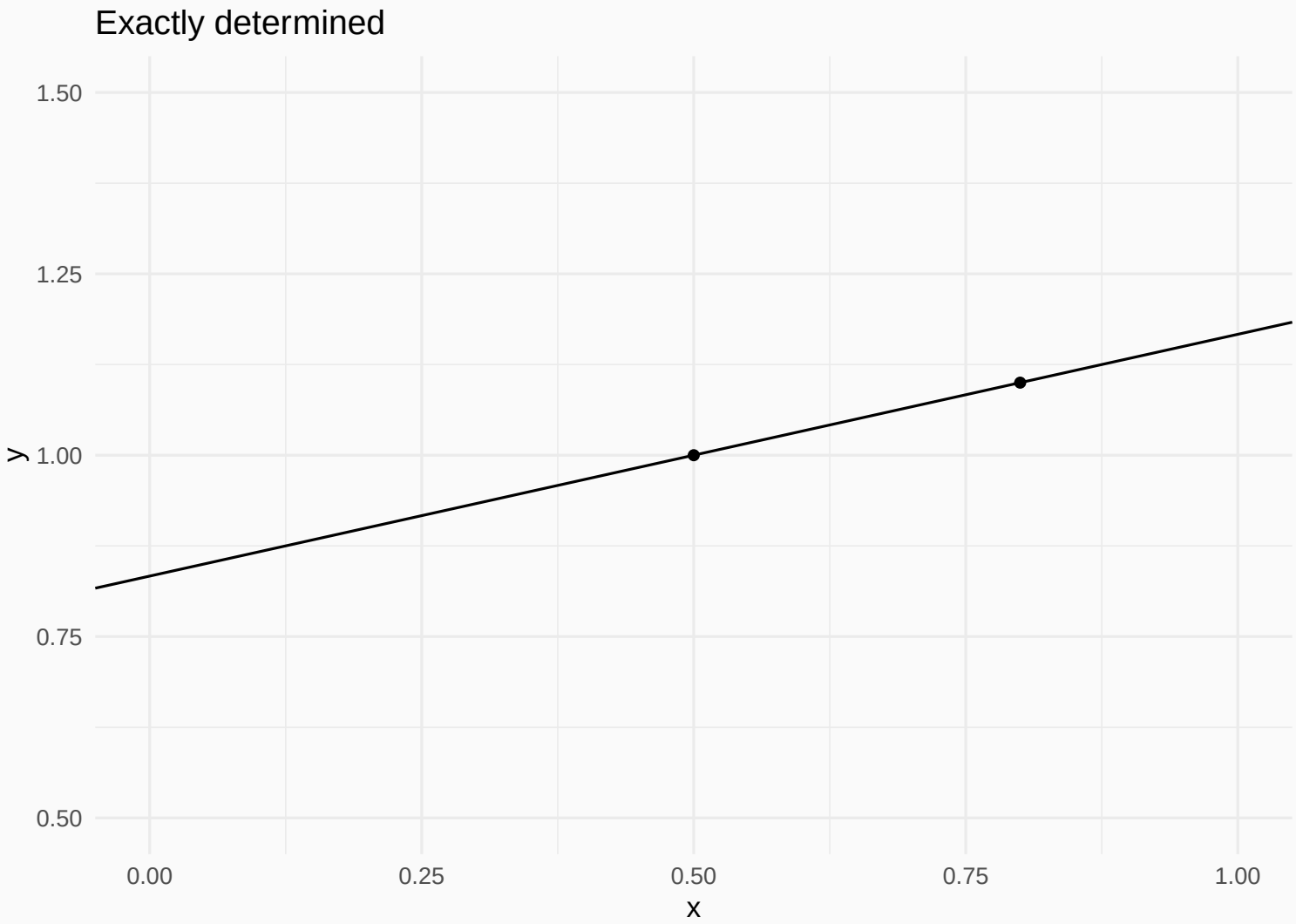
The curse of dimensionality

```
## (Intercept)           x
##   0.6578947   0.5789474
```



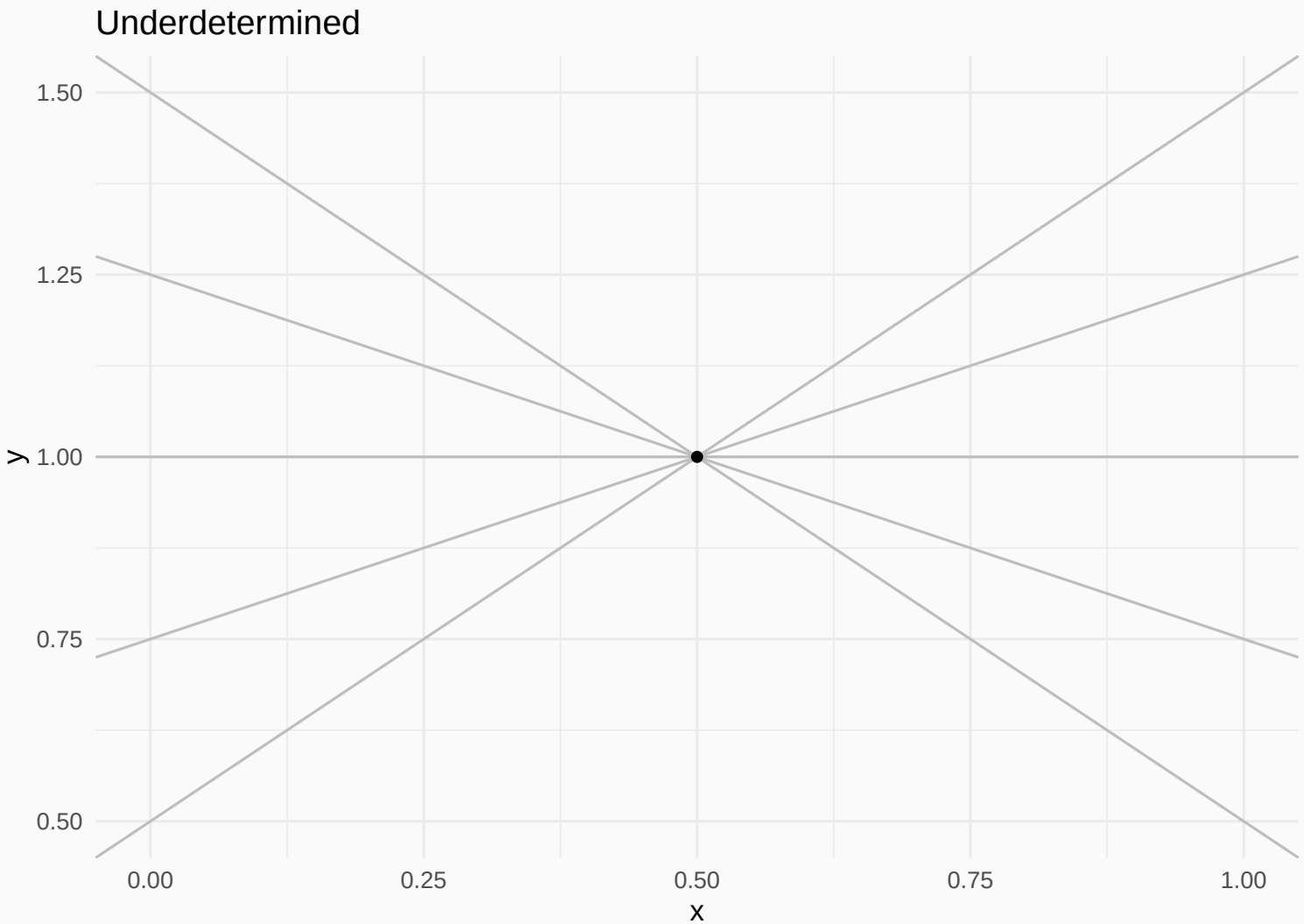
The curse of dimensionality

```
## (Intercept)           x
##   0.8333333   0.3333333
```



The curse of dimensionality

```
## (Intercept)           x
##           1          NA
```



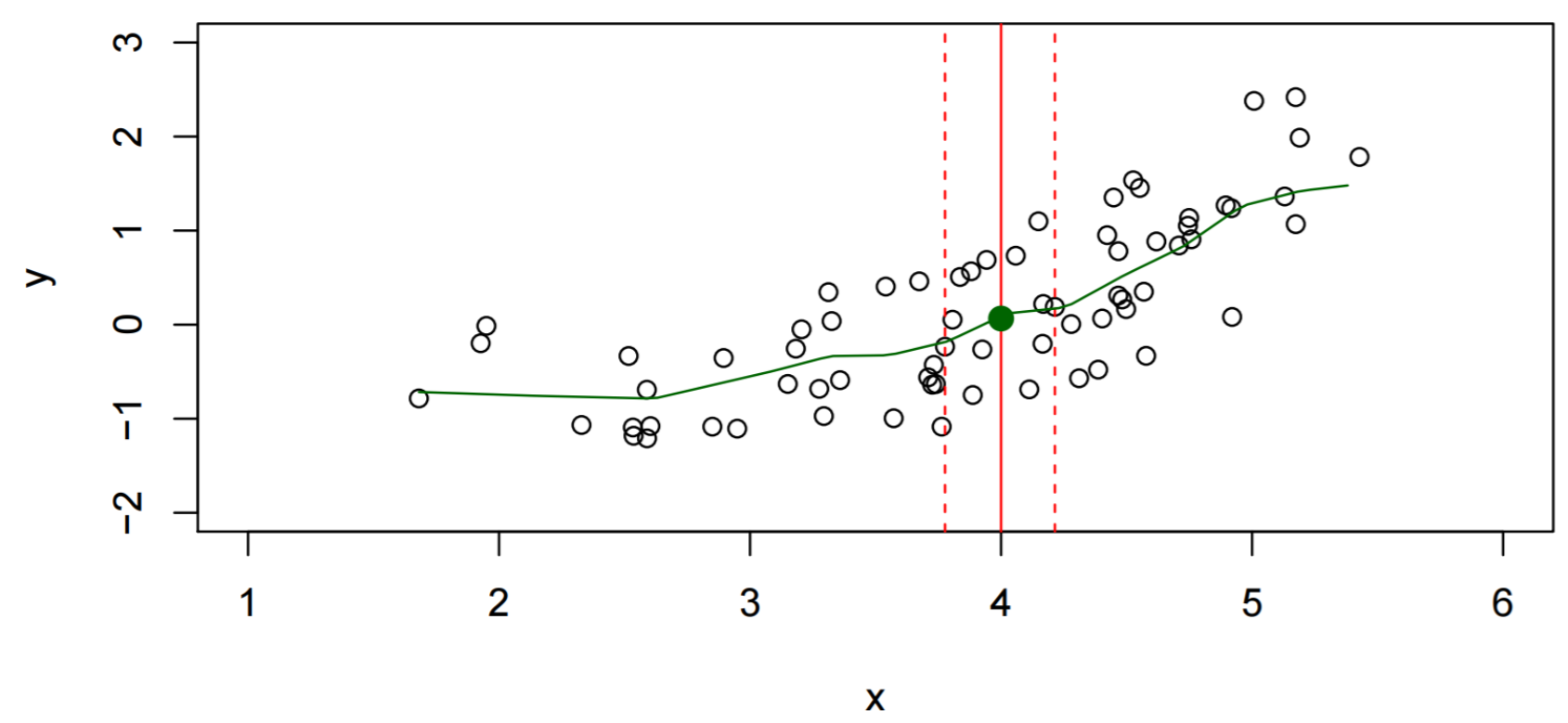
The curse of dimensionality

- In general, with k pieces of information (data points), we can (just) estimate p parameters.

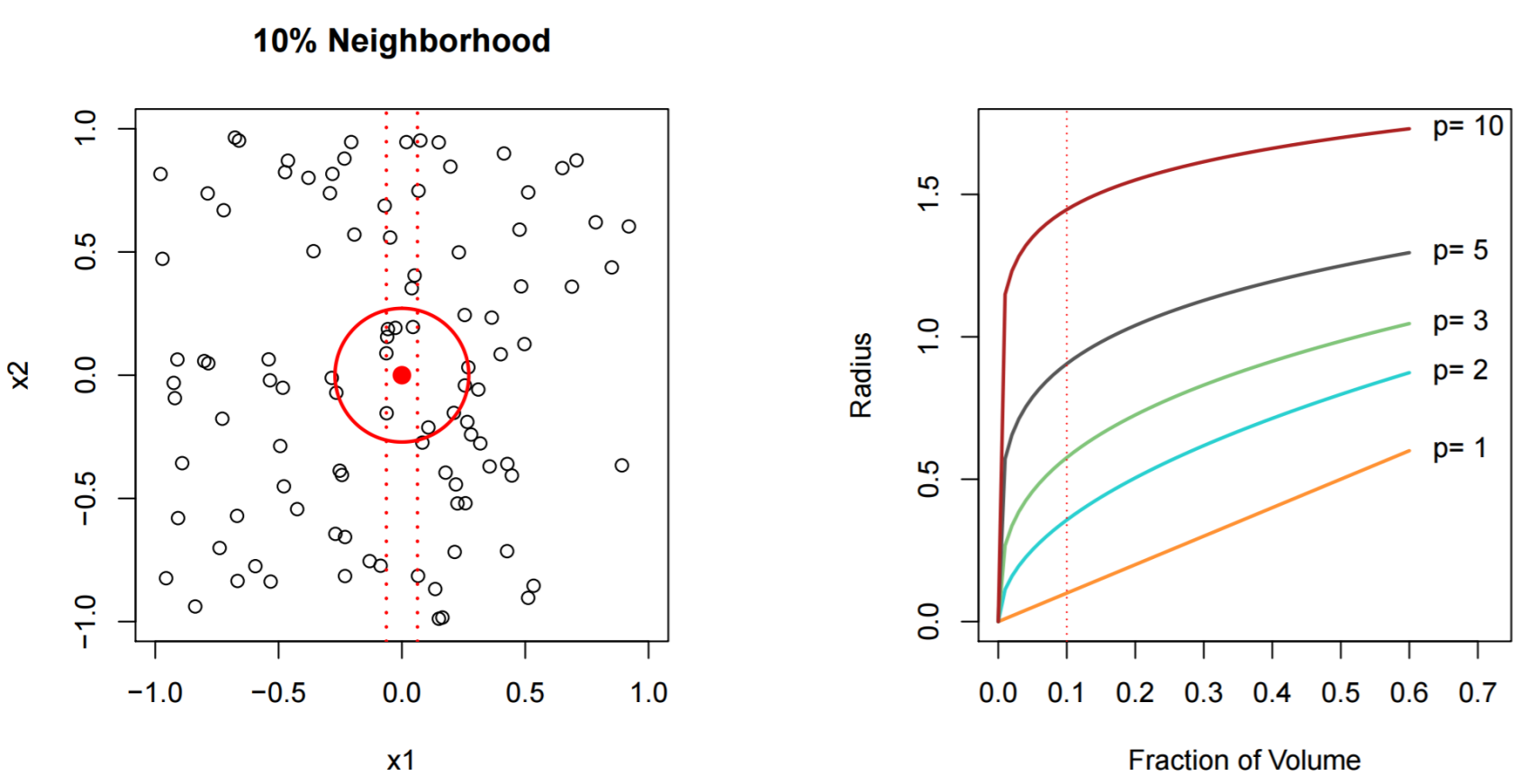
The curse of dimensionality, part II

Remember KNN regression?

$$\hat{y}_i = \frac{1}{K} \sum_{j \in J} y_j$$



The curse of dimensionality, part II



In 10 dimensions, to get 10% of uniformly distributed data we must cover 80% of the range of each input variable (ESL, par. 2.5)

The curse of dimensionality, part II

Generated data: 1 variable predicts the outcome. Now we add random noise variables. How does KNN perform? (ISLR, p. 110)

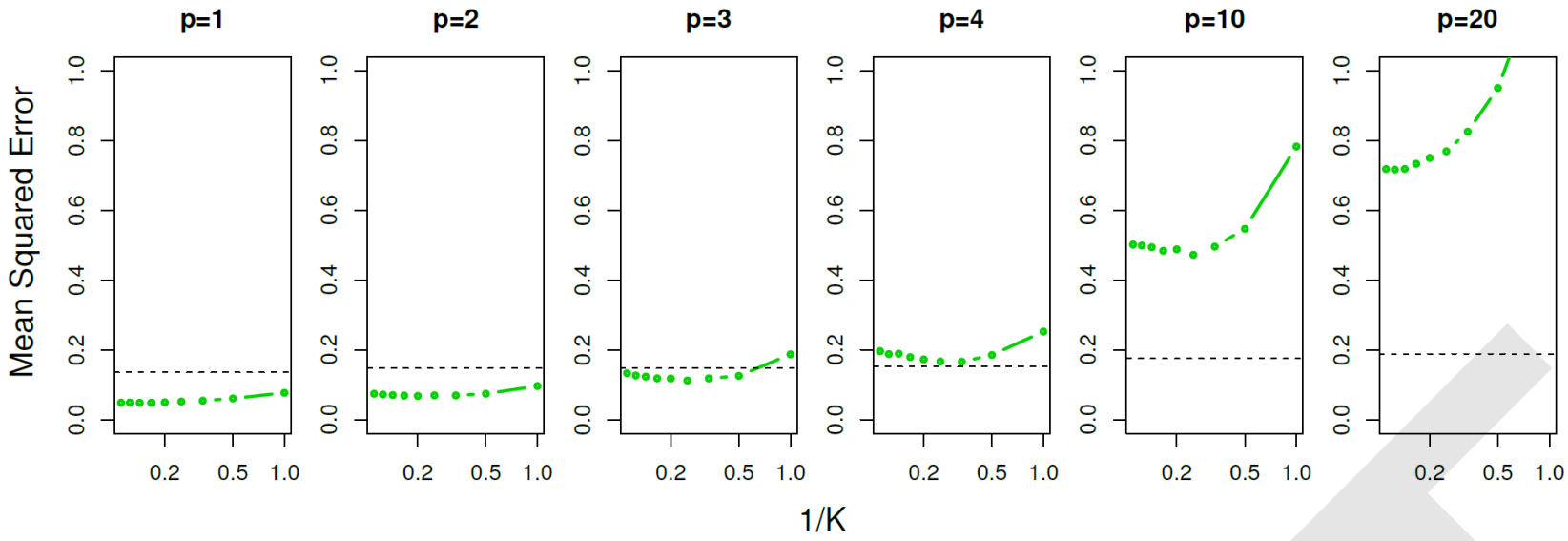


FIGURE 3.20. *Test MSE for linear regression (black dashed lines) and KNN*

The curse of dimensionality,

- If we want to estimate p parameters with a sample size of $N << p$, we need to make additional assumptions